Decomposition methods in the social sciences

Bamberg Graduate School of Social Sciences, June 7–8, 2018

Ben Jann

University of Bern, Institut of Sociology

The transformation problem

Some issues with the Oaxaca-Blinder decomposition

- The OB decomposition seems useful and easy to understand, but there are several complications we need to discuss.
  - The index problem
  - **The transformation problem / base category problem**
  - Functional form
  - Self-selection and endogeneity (not covered in this course)

Contents

1. Detailed decomposition: some conceptual complications
2. Transformation of covariates
3. Base level of categorical covariates
4. Solutions to the base level problem

Detailed decomposition: some conceptual complications

- A problem with the detailed decomposition of the "unexplained" part Δ_S^μ of the OB decomposition is that it is not invariant against (uninformative) transformations of the covariates (X variables).
- Furthermore, for categorical covariates, the results of the detailed decomposition of Δ_S^μ depend on the choice of the base/reference category.
- Some authors speak of an "identification" problem in this context. As argued by Fortin et al. (2011), however, it is more a conceptual problem of interpretation.
- The detailed decomposition of the "explained" part Δ_X^μ is more robust against these problems. Here only the contributions of the single categories of a categorical variable depend on the choice of the base category, but the sum across categories is not affected. Likewise, uninformative transformations of continuous covariates do not change the results of the detailed decomposition of Δ_X^μ.

1 Detailed decomposition: some conceptual complications

2 Transformation of covariates

3 Base level of categorical covariates

4 Solutions to the base level problem

Transformation of covariates

- Assume that a location shift (e.g. mean centering) is applied to variable X_k, that is,

$$\tilde{X}_k = X_k + \gamma$$

- Consequences of the transformation:
  - Change in the expected value of the variable:

$$\mathrm{E}(\tilde{X}_k) = \mathrm{E}(X_k + \gamma) = \mathrm{E}(X_k) + \gamma$$

  - The slope parameter β_k of the variable in a regression model is not affected, that is, $\tilde{\beta}_k = \beta_k$. Likewise, all other slope parameters are unaffected.
  - However, the intercept β_0 changes:

$$\begin{aligned} E(Y) &= \beta_0 + \beta_k E(X_k) + \textstyle\sum_{j \neq k} \beta_j E(X_j) \\ \Rightarrow \quad \beta_0 &= \mathrm{E}(Y) - \beta_k E(X_k) - \textstyle\sum_{j \neq k} \beta_j E(X_j) \\ \Rightarrow \quad \tilde{\beta}_0 &= \mathrm{E}(Y) - \beta_k (E(X_k) + \gamma) - \textstyle\sum_{j \neq k} \beta_j E(X_j) \\ &= \mathrm{E}(Y) - \beta_k E(X_k) - \textstyle\sum_{j \neq k} \beta_j E(X_j) - \beta_k \gamma = \beta_0 - \beta_k \gamma \end{aligned}$$

Transformation of covariates

- How does this affect the detailed decomposition results?
- There is no problem for the detailed decomposition of the "explained" part (as long as the same transformation is applied in both groups):

$$\begin{aligned}
\Delta^{\mu}_{X,\tilde{X}_k} &= \tilde{\beta}^0_k(\mathsf{E}(\tilde{X}_k|G=0) - \mathsf{E}(\tilde{X}_k|G=1)) \\
&= \beta^0_k(\mathsf{E}(X_k|G=0) + \gamma - \mathsf{E}(X_k|G=1) - \gamma) \\
&= \beta^0_k(\mathsf{E}(X_k|G=0) - \mathsf{E}(X_k|G=1)) \\
&= \Delta^{\mu}_{X,X_k}
\end{aligned}$$

Transformation of covariates

- The detailed decomposition of the unexplained part, however, may change:

$$
\begin{aligned}
\Delta^{\mu}_{S,\tilde{\beta}_0} &= (\tilde{\beta}^0_0 - \tilde{\beta}^1_0) = ((\beta^0_0 - \beta^0_k\gamma) - (\beta^1_0 - \beta^1_k\gamma)) \\
&= (\beta^0_0 - \beta^1_0) - \gamma(\beta^0_k - \beta^1_k) \\
&\neq (\beta^0_0 - \beta^1_0) = \Delta^{\mu}_{S,\beta_0} \\
\Delta^{\mu}_{S,\tilde{\beta}_k} &= (\tilde{\beta}^0_k - \tilde{\beta}^1_k)\,\mathrm{E}(\tilde{X}_k|G=1) \\
&= (\beta^0_k - \beta^1_k)(\mathrm{E}(X_k|G=1) + \gamma) \\
&= (\beta^0_k - \beta^1_k)\,\mathrm{E}(X_k|G=1) + \gamma(\beta^0_k - \beta^1_k) \\
&\neq (\beta^0_k - \beta^1_k)\,\mathrm{E}(X_k|G=1) = \Delta^{\mu}_{S,\beta_k}
\end{aligned}
$$

Example

```
. use gsoep29, clear
(BCPGEN: Nov 12, 2013 17:15:52-251 DBV29)
. keep if inrange(2012 - bcgeburt, 25, 55)
(10,780 observations deleted)
. generate lnwage = ln(labgro12 / (bctatzeit * 4.3)) if labgro12>0 & bctatzeit>0
(1,936 missing values generated)
. generate schooling = bcbilzeit if bcbilzeit>0
(318 missing values generated)
. generate ft_experience = expft12 if expft12>=0
(15 missing values generated)
. generate ft_experience2 = expft12^2 if expft12>=0
(15 missing values generated)
. summarize schooling if !missing(lnwage, schooling, ft_experience, ft_experience2)
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   schooling |      7,860    12.93117    2.733759         7         18
. generate c_schooling = schooling - r(mean)
(318 missing values generated)
```

Example

```
. oaxaca lnwage schooling (experience: ft_exp*), by(bcsex) weight(1)
Blinder-Oaxaca decomposition                    Number of obs     =      7,860
                                                  Model           =     linear
Group 1: bcsex = 1                                N of obs 1      =       3877
Group 2: bcsex = 2                                N of obs 2      =       3983
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	2.749054	.009236	297.64	0.000	2.730951	2.767156
group_2	2.498484	.0092013	271.54	0.000	2.48045	2.516518
difference	.2505696	.0130372	19.22	0.000	.2250172	.276122
explained	.1492473	.009391	15.89	0.000	.1308412	.1676533
unexplained	.1013223	.0131188	7.72	0.000	.07561	.1270346
explained						
schooling	-.008201	.0057638	-1.42	0.155	-.0194978	.0030958
experience	.1574483	.0080355	19.59	0.000	.1416989	.1731976
unexplained						
schooling	.0852652	.0554512	1.54	0.124	-.0234172	.1939476
experience	.1276546	.0245238	5.21	0.000	.0795889	.1757204
_cons	-.1115975	.0658889	-1.69	0.090	-.2407374	.0175423

```
experience: ft_experience ft_experience2
```

Example

```
. oaxaca lnwage c_schooling (experience: ft_exp*), by(bcsex) weight(1)
Blinder-Oaxaca decomposition                    Number of obs     =      7,860
                                                  Model           =     linear
Group 1: bcsex = 1                                N of obs 1      =       3877
Group 2: bcsex = 2                                N of obs 2      =       3983
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	2.749054	.009236	297.64	0.000	2.730951	2.767156
group_2	2.498484	.0092013	271.54	0.000	2.48045	2.516518
difference	.2505696	.0130372	19.22	0.000	.2250172	.276122
explained	.1492473	.009391	15.89	0.000	.1308412	.1676533
unexplained	.1013223	.0131188	7.72	0.000	.07561	.1270346
explained						
c_schooling	-.008201	.0057638	-1.42	0.155	-.0194978	.0030958
experience	.1574483	.0080355	19.59	0.000	.1416989	.1731976
unexplained						
c_schooling	.0002849	.000336	0.85	0.397	-.0003737	.0009434
experience	.1276546	.0245238	5.21	0.000	.0795889	.1757204
_cons	-.0266172	.0308043	-0.86	0.388	-.0869925	.0337582

```
experience: ft_experience ft_experience2
```

1 Detailed decomposition: some conceptual complications

2 Transformation of covariates

3 Base level of categorical covariates

4 Solutions to the base level problem

Base level of categorical covariates

- Changing the base level of a categorical covariate has consequences both for the detailed decomposition of Δ_X^μ and Δ_S^μ.
- Let d_j, $j = 1, \ldots, J$, be a set of indicator variables coding a categorical variable D that has J levels ($d_j = 1$ if $D = j$ and else 0).
- The contribution of D to Δ_X^μ then is:

$$\Delta_{X,D}^\mu = \beta_{d_1}^0(\bar{d}_1^0 - \bar{d}_1^1) + \beta_{d_2}^0(\bar{d}_2^0 - \bar{d}_2^1) + \cdots + \beta_{d_J}^0(\bar{d}_J^0 - \bar{d}_J^1)$$

- To estimate the coefficients, one of the levels has to be omitted (the base level). This is equivalent to constraining its coefficient to be zero.
- That is, what we are estimating are coefficients

$$\beta_{d_j^o} = \beta_{d_j} - \beta_{d_o}$$

Base level of categorical covariates

- If we omit the first level, we have

$$\Delta^{\mu}_{X,D} = 0(\bar{d}^0_1 - \bar{d}^1_1) + \beta^0_{d^1_2}(\bar{d}^0_2 - \bar{d}^1_2) + \cdots + \beta^0_{d^1_J}(\bar{d}^0_J - \bar{d}^1_J)$$

- If we omit the second level, we have

$$\Delta^{\mu}_{X,D} = \beta^0_{d^2_1}(\bar{d}^0_1 - \bar{d}^1_1) + 0(\bar{d}^0_2 - \bar{d}^1_2) + \cdots + \beta^0_{d^2_J}(\bar{d}^0_J - \bar{d}^1_J)$$

- We clearly see that individual contributions of the single indicators variables will be different depending on the choice of the base level.

Base level of categorical covariates

- However, the sum across the contributions of all indicators will always be the same. Because $\bar{d}_o = 1 - \sum_{j \neq o} \bar{d}_j$ and $\beta_{d_j^o} = \beta_{d_j} - \beta_{d_o}$ we have, for example,

$$
\begin{aligned}
\Delta^{\mu}_{X,D} &= \beta^0_{d_2^1}(\bar{d}^0_2 - \bar{d}^1_2) + \cdots + \beta^0_{d_J^1}(\bar{d}^0_J - \bar{d}^1_J) \\
&= (\beta^0_{d_2} - \beta^0_{d_1})(\bar{d}^0_2 - \bar{d}^1_2) + \cdots + (\beta^0_{d_J} - \beta^0_{d_1})(\bar{d}^0_J - \bar{d}^1_J) \\
&= \beta^0_{d_2}(\bar{d}^0_2 - \bar{d}^1_2) + \cdots + \beta^0_{d_J}(\bar{d}^0_J - \bar{d}^1_J) \\
&\qquad - \beta^0_{d_1}(\bar{d}^0_2 - \bar{d}^1_2 + \cdots + \bar{d}^0_J - \bar{d}^1_J) \\
&= \beta^0_{d_2}(\bar{d}^0_2 - \bar{d}^1_2) + \cdots + \beta^0_{d_J}(\bar{d}^0_J - \bar{d}^1_J) \\
&\qquad - \beta^0_{d_1}\{(1 - \bar{d}^0_1) - (1 - \bar{d}^2_1)\} \\
&= \beta^0_{d_1}(\bar{d}^0_1 - \bar{d}^1_1) + \beta^0_{d_2}(\bar{d}^0_2 - \bar{d}^1_2) + \cdots + \beta^0_{d_J}(\bar{d}^0_J - \bar{d}^1_J)
\end{aligned}
$$

- That is, independent of the choice of the base level, we always get the same expression.

Base level of categorical covariates

- Now consider the effect on the contributions to Δ_S^μ. Omitting the first indicator, we have

$$\Delta_S^\mu = (\beta_0^0 - \beta_0^1) + (\beta_{d_2^1}^0 - \beta_{d_2^1}^1)\bar{d}_1^1 + \cdots + (\beta_{d_J^1}^0 - \beta_{d_J^1}^1)\bar{d}_J^1 + \ldots$$

- Changing the base level has consequences for the estimated coefficients of the dummy variables (see above), but it also affects the intercept β_0.
- The intercept is equal to the expectation of Y given all covariates are zero. All $(J-1)$ included indicators being zero implies that the omitted category applies. That is, the intercept reflects the conditional outcome in the base category.
- Hence, the difference in intercepts between the two groups, $\beta_0^0 - \beta_0^1$, refers to the difference in conditional outcomes in the base category. Changing the base category changes the meaning of $\beta_0^0 - \beta_0^1$.
- Naturally, also the contributions of single indicators – as well as the sum over the contributions of all included indicators – will change.

Example

```
. recode casmin12 (1 2 = 1 "low") (4 6 = 2 "medium general") ///
>     (3 5 7 = 3 "medium vocational") (8 9 = 4 "high") (else = .) ///
>     , into(casmin4)
(8013 differences between casmin12 and casmin4)
. tab casmin4, gen(casmin4_)
```

RECODE of casmin12 (CASMIN-Klassifikation)	Freq.	Percent	Cum.
low	662	6.78	6.78
medium general	562	5.75	12.53
medium vocational	6,038	61.82	74.35
high	2,505	25.65	100.00
Total	9,767	100.00	

```
. su casmin4_*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
casmin4_1	9,767	.0677793	.2513796	0	1
casmin4_2	9,767	.0575407	.2328848	0	1
casmin4_3	9,767	.6182042	.4858518	0	1
casmin4_4	9,767	.2564759	.4367099	0	1

Example

```
. oaxaca lnwage casmin4_2 casmin4_3 casmin4_4 (experience: ft_exp*), by(bcsex) weight(1)
Blinder-Oaxaca decomposition                    Number of obs     =      7,908
                                                  Model           =     linear
Group 1: bcsex = 1                                N of obs 1      =       3898
Group 2: bcsex = 2                                N of obs 2      =       4010
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	2.748546	.0092193	298.13	0.000	2.730476	2.766615
group_2	2.498742	.0091681	272.55	0.000	2.480772	2.516711
difference	.2498043	.0130019	19.21	0.000	.2243211	.2752876
explained	.1439862	.0093367	15.42	0.000	.1256865	.1622858
unexplained	.1058182	.0134022	7.90	0.000	.0795503	.132086
explained						
casmin4_2	.0002573	.001125	0.23	0.819	-.0019478	.0024623
casmin4_3	-.0002933	.0031518	-0.09	0.926	-.0064708	.0058842
casmin4_4	.0052819	.0078272	0.67	0.500	-.0100592	.0206229
experience	.1387403	.0078625	17.65	0.000	.1233301	.1541506
unexplained						
casmin4_2	.0052416	.003508	1.49	0.135	-.001634	.0121172
casmin4_3	.0240202	.0332685	0.72	0.470	-.0411848	.0892252
casmin4_4	.0331881	.0151632	2.19	0.029	.0034688	.0629075
experience	.1277667	.0251396	5.08	0.000	.0784941	.1770394
_cons	-.0843985	.0577382	-1.46	0.144	-.1975633	.0287664

Example

```
. oaxaca lnwage casmin4_1 casmin4_2 casmin4_4 (experience: ft_exp*), by(bcsex) weight(1)
Blinder-Oaxaca decomposition                    Number of obs     =      7,908
                                                  Model           =     linear
Group 1: bcsex = 1                                N of obs 1      =       3898
Group 2: bcsex = 2                                N of obs 2      =       4010
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	2.748546	.0092193	298.13	0.000	2.730476	2.766615
group_2	2.498742	.0091681	272.55	0.000	2.480772	2.516711
difference	.2498043	.0130019	19.21	0.000	.2243211	.2752876
explained	.1439862	.0093367	15.42	0.000	.1256865	.1622858
unexplained	.1058182	.0134022	7.90	0.000	.0795503	.132086
explained						
casmin4_1	.0019874	.0014702	1.35	0.176	-.0008941	.004869
casmin4_2	-.0000604	.0002672	-0.23	0.821	-.000584	.0004633
casmin4_4	.0033187	.0049168	0.67	0.500	-.006318	.0129555
experience	.1387403	.0078625	17.65	0.000	.1233301	.1541506
unexplained						
casmin4_1	-.0021358	.0029612	-0.72	0.471	-.0079397	.0036681
casmin4_2	.0034315	.0026602	1.29	0.197	-.0017824	.0086453
casmin4_4	.0227487	.0073685	3.09	0.002	.0083067	.0371907
experience	.1277667	.0251396	5.08	0.000	.0784941	.1770394
_cons	-.045993	.0342941	-1.34	0.180	-.1132081	.0212222

Example

```
. oaxaca lnwage (casmin: casmin4_2 casmin4_3 casmin4_4) (experience: ft_exp*), ///
>     by(bcsex) weight(1)
Blinder-Oaxaca decomposition                    Number of obs     =      7,908
                                                  Model           =     linear
Group 1: bcsex = 1                                N of obs 1      =       3898
Group 2: bcsex = 2                                N of obs 2      =       4010
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	2.748546	.0092193	298.13	0.000	2.730476	2.766615
group_2	2.498742	.0091681	272.55	0.000	2.480772	2.516711
difference	.2498043	.0130019	19.21	0.000	.2243211	.2752876
explained	.1439862	.0093367	15.42	0.000	.1256865	.1622858
unexplained	.1058182	.0134022	7.90	0.000	.0795503	.132086
explained						
casmin	.0052458	.0053637	0.98	0.328	-.0052669	.0157585
experience	.1387403	.0078625	17.65	0.000	.1233301	.1541506
unexplained						
casmin	.0624499	.0495092	1.26	0.207	-.0345864	.1594862
experience	.1277667	.0251396	5.08	0.000	.0784941	.1770394
_cons	-.0843985	.0577382	-1.46	0.144	-.1975633	.0287664

```
casmin: casmin4_2 casmin4_3 casmin4_4
experience: ft_experience ft_experience2
```

Example

```
. oaxaca lnwage (casmin: casmin4_1 casmin4_2 casmin4_4) (experience: ft_exp*), ///
>     by(bcsex) weight(1)
Blinder-Oaxaca decomposition                    Number of obs     =      7,908
                                                  Model           =     linear
Group 1: bcsex = 1                                N of obs 1      =       3898
Group 2: bcsex = 2                                N of obs 2      =       4010
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	2.748546	.0092193	298.13	0.000	2.730476	2.766615
group_2	2.498742	.0091681	272.55	0.000	2.480772	2.516711
difference	.2498043	.0130019	19.21	0.000	.2243211	.2752876
explained	.1439862	.0093367	15.42	0.000	.1256865	.1622858
unexplained	.1058182	.0134022	7.90	0.000	.0795503	.132086
explained						
casmin	.0052458	.0053637	0.98	0.328	-.0052669	.0157585
experience	.1387403	.0078625	17.65	0.000	.1233301	.1541506
unexplained						
casmin	.0240444	.009222	2.61	0.009	.0059695	.0421193
experience	.1277667	.0251396	5.08	0.000	.0784941	.1770394
_cons	-.045993	.0342941	-1.34	0.180	-.1132081	.0212222

```
casmin: casmin4_1 casmin4_2 casmin4_4
experience: ft_experience ft_experience2
```

1 Detailed decomposition: some conceptual complications

2 Transformation of covariates

3 Base level of categorical covariates

4 Solutions to the base level problem

Normalization

- „Normalization“ of the coefficients associated to with categorical variables has been suggested as a solution to the problem that the choice of the base level changes the detailed decomposition results.
- One solution are so-called “deviation contrasts” (equivalent to “effect coding”): the coefficients of the indicators reflect deviations from the unweighted average across categories (balanced grand mean).
- The decomposition results based on coefficients that have been normalized using the deviation contrast transform are independent of the choice of the base level.
- Furthermore, the results are equal to the (unweighted) average of the results one would get from a series of decompositions in which the categories are used one after another as the base level (Yun 2008).

Normalization

- The deviation contrast normalization works as follows:
  - Again, let d_j, $j = 1, \ldots, J$, be a set of indicator variables and $\beta_{d_j^o}$ be the corresponding coefficients (with $\beta_{d_o^o} = 0$).
  - Determine
$$c = \frac{\beta_{d_1^o} + \cdots + \beta_{d_J^o}}{J}$$
  - Compute the transformed coefficients
$$\tilde{\beta}_0 = \beta_0 + c \quad \text{and} \quad \beta_{d_j} = \beta_{d_j^o} - c$$
(note that $\sum_{j=1}^{J} \beta_{d_j} = 0$).
  - Use coefficients $\tilde{\beta}_0$ and β_{d_j} to perform the decomposition instead of the original coefficients.
  - An alternative to transforming the coefficients would be to apply restricted least-squares estimation with restriction $\sum_{j=1}^{J} \beta_{d_j} = 0$.

Example

```
. regress lnwage casmin4_2 casmin4_3 casmin4_4 ft_exp*, noheader
```

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
casmin4_2	.1761139	.0372611	4.73	0.000	.1030724	.2491554
casmin4_3	.2642998	.0268254	9.85	0.000	.2117149	.3168848
casmin4_4	.7138012	.0280756	25.42	0.000	.6587656	.7688368
ft_experience	.042032	.0021529	19.52	0.000	.0378117	.0462524
ft_experience2	-.0007705	.0000638	-12.07	0.000	-.0008956	-.0006453
_cons	1.876787	.0283324	66.24	0.000	1.821248	1.932326

```
. devcon, groups(casmin4_1 casmin4_2 casmin4_3 casmin4_4)
Transformed regress coefficients                Number of obs    =      7908
```

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
casmin4_1	-.2885537	.020688	-13.95	0.000	-.3291076	-.2479998
casmin4_2	-.1124398	.0215687	-5.21	0.000	-.1547203	-.0701594
casmin4_3	-.0242539	.0113812	-2.13	0.033	-.046564	-.0019438
casmin4_4	.4252475	.0127094	33.46	0.000	.4003336	.4501613
ft_experience	.042032	.0021529	19.52	0.000	.0378117	.0462524
ft_experience2	-.0007705	.0000638	-12.07	0.000	-.0008956	-.0006453
_cons	2.165341	.0157012	137.91	0.000	2.134562	2.196119

Example

```
. regress lnwage casmin4_1 casmin4_2 casmin4_4 ft_exp*, noheader
```

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
casmin4_1	-.2642998	.0268254	-9.85	0.000	-.3168848	-.2117149
casmin4_2	-.0881859	.0282609	-3.12	0.002	-.1435849	-.032787
casmin4_4	.4495014	.0135685	33.13	0.000	.4229036	.4760992
ft_experience	.042032	.0021529	19.52	0.000	.0378117	.0462524
ft_experience2	-.0007705	.0000638	-12.07	0.000	-.0008956	-.0006453
_cons	2.141087	.0160216	133.64	0.000	2.10968	2.172493

```
. devcon, groups(casmin4_1 casmin4_2 casmin4_3 casmin4_4)
Transformed regress coefficients                Number of obs    =      7908
```

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
casmin4_1	-.2885537	.020688	-13.95	0.000	-.3291076	-.2479998
casmin4_2	-.1124398	.0215687	-5.21	0.000	-.1547203	-.0701594
casmin4_3	-.0242539	.0113812	-2.13	0.033	-.046564	-.0019438
casmin4_4	.4252475	.0127094	33.46	0.000	.4003336	.4501613
ft_experience	.042032	.0021529	19.52	0.000	.0378117	.0462524
ft_experience2	-.0007705	.0000638	-12.07	0.000	-.0008956	-.0006453
_cons	2.165341	.0157012	137.91	0.000	2.134562	2.196119

Example

```
. oaxaca lnwage normalize(casmin4_*) (experience: ft_exp*), by(bcsex) weight(1)
(normalized: casmin4_1 casmin4_2 casmin4_3 casmin4_4)
Blinder-Oaxaca decomposition                    Number of obs     =      7,908
                                                  Model           =     linear
Group 1: bcsex = 1                                N of obs 1      =       3898
Group 2: bcsex = 2                                N of obs 2      =       4010
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	2.748546	.0092193	298.13	0.000	2.730476	2.766615
group_2	2.498742	.0091681	272.55	0.000	2.480772	2.516711
difference	.2498043	.0130019	19.21	0.000	.2243211	.2752876
explained	.1439862	.0093367	15.42	0.000	.1256865	.1622858
unexplained	.1058182	.0134022	7.90	0.000	.0795503	.132086
explained						
casmin4_1	.0022361	.0016403	1.36	0.173	-.0009788	.005451
casmin4_2	-.0001001	.0004385	-0.23	0.819	-.0009595	.0007593
casmin4_3	.0000367	.0003947	0.09	0.926	-.0007368	.0008102
casmin4_4	.0030731	.0045529	0.67	0.500	-.0058504	.0119967
experience	.1387403	.0078625	17.65	0.000	.1233301	.1541506
unexplained						
casmin4_1	-.0037775	.0022963	-1.65	0.100	-.0082782	.0007231
casmin4_2	.00204	.002024	1.01	0.313	-.0019269	.006007
casmin4_3	-.0184644	.0141663	-1.30	0.192	-.0462298	.0093009
casmin4_4	.0147239	.0068526	2.15	0.032	.001293	.0281548
experience	.1277667	.0251396	5.08	0.000	.0784941	.1770394
_cons	-.0164705	.0330376	-0.50	0.618	-.081223	.048282

```
experience: ft_experience ft_experience2
```

Example

```
. forv j = 1(1)4 {
  2.     local casmin casmin4_1 casmin4_2 casmin4_3 casmin4_4 _cons
  3.     local casmin: subinstr local casmin "casmin4_`j'" ""
  4.     quietly oaxaca lnwage `casmin' (experience: ft_exp*), by(bcsex) weight(1)
  5.     estimates store m`j'
  6. }
. estout m*, keep(unexplained:casmin4_* unexplained:_cons) order(casmin4_1) collab(none)
```

	m1	m2	m3	m4
unexplained				
casmin4_1		-.0061845	-.0021358	-.0067899
casmin4_2	.0052416		.0034315	-.000513
casmin4_3	.0240202	-.045535		-.0523429
casmin4_4	.0331881	.0029588	.0227487	
_cons	-.0843985	.0268122	-.045993	.0376973

```
. mata: mean(editmissing(st_matrix("r(coefs)"), 0)')'
                 1
  +-----------------+
1 |  -.0037775412   |
2 |   .0020400073   |
3 |  -.0184644336   |
4 |   .0147239074   |
5 |  -.0164704983   |
  +-----------------+
```

Normalization

- An alternative – and probably superior – variant of the normalization uses coefficients that reflect deviations from the weighted average across categories (observation-weighted grand mean), where the weights are proportional to the probabilities of the categories (Kennedy 1986, Haisken-DeNew and Schmidt 1997).
- That is, use

$$c = \Pr(D = 1)\beta_{d_1^o} + \cdots + \Pr(D = J)\beta_{d_J^o}$$

such that

$$\sum_{j=1}^{J} \Pr(D = 1)\beta_{d_j} = 0$$

- This limits the influence of sparsely populated categories and makes results more robust against recoding the categorical variable (i.e. combining several sparsely populated categories into one will not have much of an effect on the results; see Kim 2013).

Normalization

- Yet another type of normalization is to compute

$$\Delta^{\mu}_{S,d_j} = (\beta^0_0 - \beta^1_0) + (\beta^0_{d^o_j} - \beta^1_{d^o_j}) + \sum_{k=1}^{K}(\beta^0_k - \beta^1_k)\bar{X}^1_k$$

  as suggested by Horrace and Oaxaca (2001), where d_j, $j = 1, \ldots, J$, is again a set of indicator variables and X_k, $k = 1, \ldots, K$ are all other covariates, and then normalize the contributions using

$$\%\Delta^{\mu}_{S,d_j} = \frac{\bar{d}^1_j \Delta^{\mu}_{S,d_j}}{\Delta^{\mu}_S} \quad \text{since } \Delta^{\mu}_S = \sum_{j=1}^{K} \bar{d}^1_j \Delta^{\mu}_{S,d_j}$$

  as suggested by Fortin et al. (2011)

- This makes sense, for example, if we want to know how much different industries contribute to the unexplained wage gap, controlling for differential composition of the industries with respect to the X variables and taking into account the industry size.

Exercise 4

- Replace ISEI in the model of exercise 1 by the (categorical) EGP variable. Only report the aggregate contribution of EGP. Illustrate how results change if you switch the base level.
- Normalize the effects of EGP to make its contribution independent of the choice of the base level.
- Now simplify the EGP variable by combining some sparsely populated categories. How do the decomposition results change?
- Compute the contribution of EGP and the simplified EGP to the unexplained part using a weighted normalization. You need to do this manually (hint: you can use command `contrast` to obtain normalized coefficients after running a regression). Compare the results to the results from the unweighted normalization.
- Generate a handful of economic sectors (e.g. primary, secondary, tertiary; possibly subdivide the tertiary sector into 2 or three subsectors) from variable `nace12`. Compute the "industry decomposition" by Horrace and Oaxaca (2001)/Fortin et al. (2011) described above.

Comment

- There is always a certain arbitrariness to the different normalization approaches. There is no right or wrong; what makes sense may depend on context.
- Fortin et al. (2011) suggest that it may be more fruitful to chose the omitted category based on substantive reasoning and stick to the original results. This requires more thinking about how the results can be meaningfully interpreted in a specific case.

References

- Fortin, Nicole, Thomas Lemieux, Sergio Firpo (2011). Decomposition Methods in Economics. Pp. 1–102 in: O. Ashenfelter and D. Card (eds.). Handbook of Labor Economics. Amsterdam: Elsevier.
- Haisken-DeNew, John P., Christoph M. Schmidt (1997). Inter-Industry and Inter-Regional Differentials: Mechanics and Interpretation. The Review of Economics and Statistics 79(3):516–521.
- Horrace, William C., Ronald L. Oaxaca (2001). Inter-Industry Wage Differentials and the Gender Wage Gap: An Identification Problem. Industrial and Labor Relations Review 54(3):611–618.
- Kennedy, Peter (1986). Interpreting Dummy Variables. The Review of Economics and Statistics 68(1):174–175.
- Kim, ChangHwan (2013). Detailed Wage Decompositions. Revisiting the Identification Problem. Sociological Methodology 43:346–363.
- Yun, Myeong-Su (2008). Identification problem and detailed Oaxaca decomposition: A general solution and statistical inference. Journal of Economic and Social Measurement 33:27–38.